IDENTIFAI

Deepfake Intelligence Report

Periodic Analysis of Synthetic Media Incidents

Author	identifAI Threat Intel Team
Period	2020 - 2026
Date	March 26, 2026
Version	1.0

Legal Disclaimer & Terms of Use

1. **Nature of the Report and Institutional Intent**

   This report, titled Deepfake Intelligence Report, is published by identifAI—a specialized entity dedicated to the development of defensive technologies against synthetic media and deepfakes. This document is intended exclusively for informational, educational, and observational purposes. It serves as a technical "Observatory" on the current state of deepfake technology and its impact on the information ecosystem. Under no circumstances shall this report be construed as a legal accusation, a formal indictment, or an attempt to defame, offend, or damage the reputation of any individual, public figure, political entity, or nation mentioned herein. No statement contained in this report is intended to allege, nor should it be construed as alleging, that any named or referenced individual has engaged in unlawful, wrongful, or otherwise reprehensible conduct. All references to named individuals – whether public figures, political representatives, corporate executives, or private persons – describe exclusively the role of such individuals as subjects of observed synthetic media incidents, meaning as targets or subjects of third-party-generated deepfake content. Such references carry no implication of personal misconduct, criminal activity, or reputational fault on the part of the named individuals.

2. **Limitation of Liability; Accuracy and Methodology**

   While identifAI strives to ensure the highest degree of accuracy and technical rigor, the field of synthetic media is characterized by rapid evolution and inherent opacity. Consequently, the information, attributions, and analyses provided are offered on an "as-is" basis, without warranties of any kind, express or implied, as to the correctness, completeness, timeliness, or fitness for any particular purpose of the data presented. All quantitative data and categorical attributions contained herein are the output of a structured incident harvesting and normalization pipeline, incorporating LLM-assisted classification subsequently reviewed and validated by human subject-matter experts. Notwithstanding such review, AI-assisted classification processes may be subject to systematic or incidental error; accordingly, no representation is made that the data is free from inaccuracy or bias. Readers are directed to Section 11 (Methodology) of this report for a full description of the analytical framework, its assumptions, and its limitations. identifAI, its affiliates, officers, employees, and contributors shall not be held liable for any direct, indirect, incidental, special, or consequential damages arising from: (i) errors or omissions in the data; (ii) the interpretation or use of findings contained in this report; or (iii) any reliance placed on this report by any third party.

3. **Non-Adversarial Attribution; Public Interest Basis**

   Any mention of specific politicians, governmental bodies, corporate entities, media organizations or nations is based on technical forensic analysis, statistical aggregation and open-source intelligence (OSINT) gathered in accordance with the methodology described in Section 11. Such mentions do not imply a judgment of intent, guilt, or culpability on the part of any mentioned party; do not constitute a claim that any named party created, disseminated, or endorsed any synthetic media content; are made solely to describe observed patterns of targeting, dissemination, and platform exploitation within the synthetic media threat landscape. The objective of this report is the analysis of technical trends and threat patterns for the purpose of advancing public awareness and organizational resilience. This report does not intervene in geopolitical discourse, partisan political activities, or any active legal proceeding. Publication on a matter of public interest: This report addresses matters of demonstrated public interest, including the integrity of democratic processes, financial security, and the protection of individuals from synthetic media abuse. The analysis and any related statements are made responsibly, on the basis of the evidence described herein, and following reasonable verification procedures consistent with responsible journalistic and research standards.

4. **AI Disclosure and Transparency**

   In accordance with identifAI's AI Disclosure Policy, and in compliance with applicable transparency obligations under Regulation (EU) 2024/1689 (EU Artificial Intelligence Act), the reader is hereby notified of the following:

   - Textual Content: The text within this report has been drafted with the assistance of Artificial Intelligence systems and subsequently reviewed and verified by human subject-matter experts to ensure technical consistency and factual accuracy. Human review constitutes the authoritative layer while AI-generated text serves as a drafting aid only.

- Visual and Synthetic Media: All images, diagrams, or synthetic examples used within this report have been generated or modified via AI-based generative systems. These visual assets are employed exclusively for illustrative and didactic purposes to demonstrate the technologies analyzed by the Observatory. They do not constitute evidentiary material, do not purport to depict real events, and, unless expressly stated otherwise, do not represent or reproduce the likeness of any real person. They are fully compliant with identifAI's internal transparency and ethical standards.

5. Privacy and Data Protection

The processing of personal data of identifiable natural persons referenced in this report, including public figures, is carried out on the following legal bases:

- Legitimate interest and public interest (Art. 6(1)(e) and (f), Regulation (EU) 2016/679 "GDPR") given the manifest public interest in monitoring and countering synthetic media threats to democratic processes and financial integrity;
- Research, journalistic, and scientific expression purposes, as recognised under Art. 85 GDPR and applicable national implementing legislation.

identifAI processes personal data to the minimum extent necessary for the analytical purposes described herein. Named individuals may exercise their rights under applicable data protection law by contacting identifAI at For readers outside the European Union: identifAI observes data protection principles consistent with GDPR standards as a baseline across all jurisdictions of distribution.

6. Intellectual Property and Permitted Use

 Reproduction, distribution, or communication of this report, in whole or in part, is permitted for non-commercial purposes only, subject to the following conditions:

- The report must be cited in full as: identifAI Deepfake Intelligence Report, [period], identifAI Threat Intel Team, [date];
- Partial extracts may not be reproduced in isolation where such extraction would distort the meaning, context or statistical findings of the original text;
- Any adaptation, translation or derivative work requires prior written consent from identifAI;
- Commercial use including incorporation into paid products, services, or reports is strictly prohibited without a separate licensing agreement.

identifAI reserves the right to take appropriate legal action against any use of this report that misrepresents its findings, decontextualizes its data or causes reputational harm to identifAI or to any party referenced herein.

Executive Summary

Definition

For the purposes of this analysis, an **incident** is strictly defined as an observed case of a deepfake being used in the real world, as reported by English-based media sources globally or through mentions of deepfakes on social media posts.

This periodic intelligence briefing provides a strategic overview of the global threat landscape based on our observations of deepfake activities spanning from January 2020 to March 2026. Across this reporting period, synthetic media has transitioned from a fringe technological novelty into a systemic vector for cognitive disruption, financial extortion, and reputational damage. Operating within these parameters, our intelligence reveals a highly industrialized ecosystem where threat actors seamlessly cross geopolitical boundaries, exploiting asymmetrical advantages to launch sophisticated campaigns. Notably, the escalation of Business Email Compromise (BEC) and advanced Tactics, Techniques, and Procedures (TTP) leveraging synthetic media presents an unprecedented challenge to enterprise integrity.

A synthesis of our global observations reveals distinct patterns across vectors, geographies, and targets. Video manipulation dominates the technical carriers at **45.6%**, directly fueling political manipulation, which constitutes **24.6%** of the total threat landscape. Fraud follows closely at **20.1%**, driven by the convergence of synthetic identities and financial scams. The United States remains the overwhelming epicenter of geographic targeting at **46.9%**, followed by the United Kingdom at **8.2%**. Concurrently, malicious actors exhibit a clear preference for exploiting algorithmic amplification on platforms like X (**51.2%**) and TikTok (**21.1%**). At the target level, while prominent global leaders bear the brunt of highly visible disinformation operations, the underlying TTPs are rapidly trickling down to exploit corporate infrastructure, blurring the lines between geopolitical subversion and transnational organized crime.

Looking toward the 12-month risk trajectory, the convergence of real-time synthetic media generation and automated deployment frameworks will likely collapse the required time-to-compromise for enterprise networks. identifAI's proactive detection capabilities remain critical in parsing this rapidly expanding volume of synthetic artifacts. As threat actors refine multi-modal biometric bypass techniques and orchestrate highly localized disinformation campaigns, institutional resilience will depend not merely on reactive mitigation, but on deeply integrated, mathematically rigorous authenticity verification at the architectural level.

Opening Report Disclaimer

Please note that as this is our opening Deepfake Intelligence Report, the observation period spans an extended timeframe (January 2020 to March 2026) to establish a comprehensive baseline. Readers should be aware that aggregating data over this six-year window can smooth over acute, short-term fluctuations, and overall percentage distributions will heavily reflect long-term structural shifts rather than just immediate current events.

Contents

1 Key Findings

45.6%
Video Deepfakes
of incidents

20.1%
Fraud Motivation
of incidents

24.6%
Political Targets
of campaigns

This periodic assessment is anchored in a comprehensive evaluation of global synthetic media events. Within our intelligence framework, an incident is an observed case of a deepfake being used in the real world, as reported by English-based media sources globally or via mentions to deepfakes on social media posts. Grounded in a baseline of about 10k total observed global incidents, the following insights strictly utilize percentage distributions to illuminate the structural dynamics of the threat environment.

Key Finding

Video formats account for **45.6%** of all observed technical carriers, with mixed media comprising **25.2%**, demonstrating a heavy reliance on multi-sensory deception to bypass human critical evaluation and automated moderation systems.

Key Finding

Political manipulation represents **24.6%** of the total threat landscape, while fraud-related incidents constitute **20.1%**, underscoring the dual monetization and destabilization motives driving advanced threat actor groups.

Key Finding

X serves as the primary distribution channel for **51.2%** of threat propagation among top social media sources, significantly outpacing TikTok (**21.1%**) and YouTube (**10.0%**) in algorithmic exposure.

Key Finding

The United States absorbs **46.9%** of the geopolitical footprint, exposing a distinct asymmetry wherein Western democratic and financial institutions remain the primary focal points for both state-aligned and financially motivated deepfake campaigns.

Deepfake Threat Landscape

2 Threat Landscape: The Deepfake Ecosystem

The synthetic media ecosystem operates as a highly segmented, purpose-driven economy. Based on our observations, the distribution of deepfake incidents reveals distinct operational priorities among threat actors. While a significant portion of incidents remains unclassified (**31.5%**) due to the rapid, ephemeral nature of certain digital artifacts, the successfully categorized events highlight a landscape dominated by political manipulation (**24.6%**) and fraud (**20.1%**). Pornography (**11.3%**), celebrity misuse (**9.3%**), and satire (**3.2%**) constitute the remainder of the primary threat vectors, each underpinned by specific systemic motivations and monetization strategies. Figure 1 illustrates the incident categories, showing the percentage share by deepfake threat type. We define as "unclassified" those incidents where the available data was insufficient to confidently attribute a specific threat category, often due to the transient nature of the content, lack of contextual information, or ambiguity in the observed patterns. For example, an unclassified incident might involve hyper-realistic, AI-generated lifestyle or entertainment videos circulating on social media that confuse younger audiences about their authenticity, yet lack any discernible political, financial, or malicious motive.

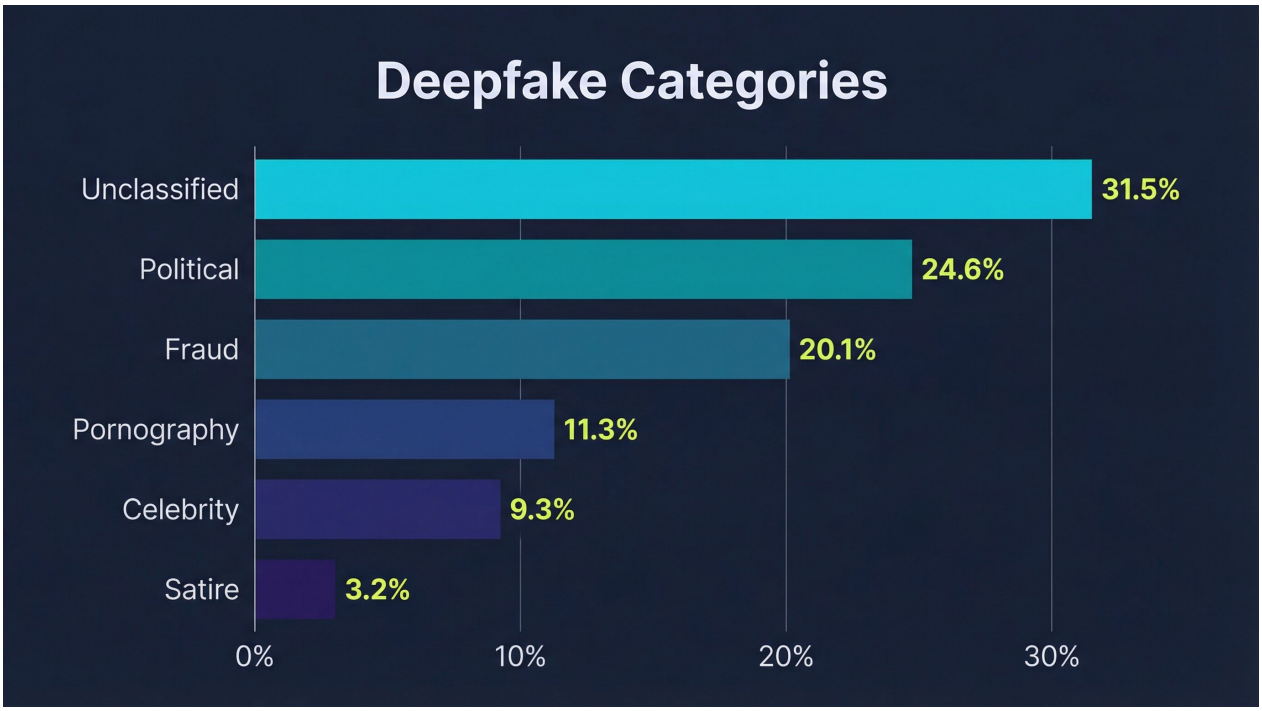


Figure 1: Incident categories – percentage share by deepfake threat type.

Political manipulation is highly effective because it deliberately leverages social engineering to weaponize existing societal polarization. Threat actors deploying these campaigns range from state-aligned advanced persistent threats (APTs) to decentralized hacktivist collectives. By targeting civic processes and public discourse, these entities aim to disrupt public trust and manufacture consensus. The efficacy of political deepfakes lies in their capacity for rapid, emotional resonance; they are engineered not necessarily to withstand forensic scrutiny, but to survive just long enough within algorithmic echo chambers to alter public perception.

Similarly, celebrity misuse relies on the inherent trust and parasocial relationships audiences have with high-profile figures. This category heavily cross-pollinates with fraud, where the unauthorized synthesis of a celebrity's likeness is utilized to endorse fraudulent investment schemes or cryptocurrency scams. Fraud itself, capturing roughly one-fifth of the ecosystem, is systematically deployed by financially motivated cybercriminal syndicates. These actors utilize synthetic media for identity theft, executing sophisticated financial scams that exploit biometric authentication workflows and corporate communication channels. The underlying systemic "why" for fraud is purely economic: deepfakes radically reduce the overhead costs of social engineering at scale, allowing attackers to mass-produce highly personalized phishing and extortion material.

Non-consensual imagery, categorized predominantly under pornography, represents a deeply pervasive vector for harassment, extortion, and reputational destruction. This vector is frequently deployed by individual malicious actors and localized extortion rings aiming to inflict psychological harm or coerce financial compliance from victims. Its persistence highlights the democratization of consumer-grade generative tools, which no longer require specialized technical acumen to operate.

Looking ahead, we forecast a compounding escalation in both political manipulation and fraud categories. As intersecting global election cycles intensify and financial extortion methodologies increasingly incorporate real-time synthetic voice and video generation, the convergence of disinformation and economic crime will aggressively stress-test existing institutional defenses. Figure 2 illustrates the deepfake carrier breakdown, showing the percentage share of total incidents by modality.

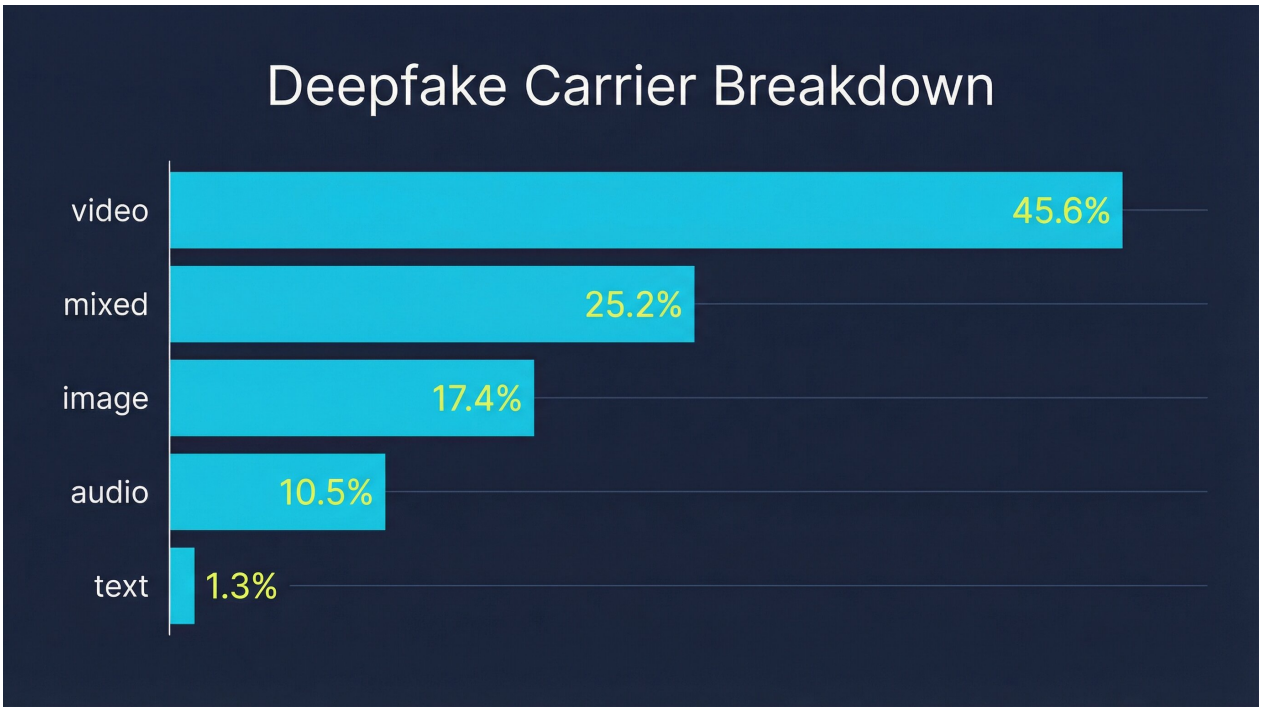
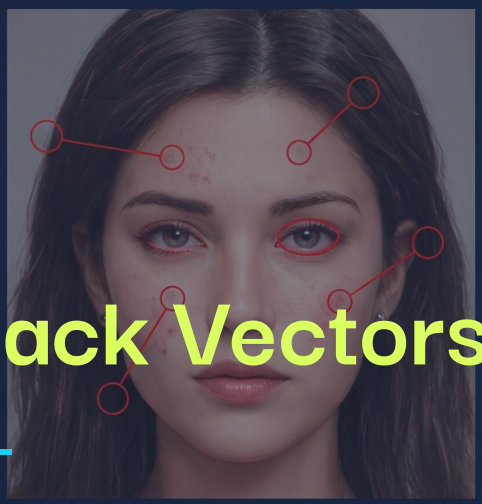


Figure 2: Deepfake carrier breakdown – percentage share of total incidents by modality.

Deepfake Modalities & Attack Vectors

VIDEO
FACE-SWAP

VOICE
CLONE



SYNTHETIC
IMAGE

3 Attack TTPs, Carriers & Technical Modalities

The technical sophistication of synthetic media operations has matured considerably throughout the reporting period from January 2020 to March 2026. A forensic evaluation of our observations maps the distribution of technical carriers, revealing a hierarchy of exploitation. AI-generated video leads the attack taxonomy at **45.6%**, followed by mixed formats (combining multiple synthetic elements) at **25.2%**, synthetic still images at **17.4%**, voice cloning at **10.5%**, and text generation at **1.3%**. This distribution directly correlates with the rapid commercialization and open-source availability of advanced neural architectures. Figure 3 depicts the monthly incident trend, illustrating the total deepfake incidents logged per month over this period.

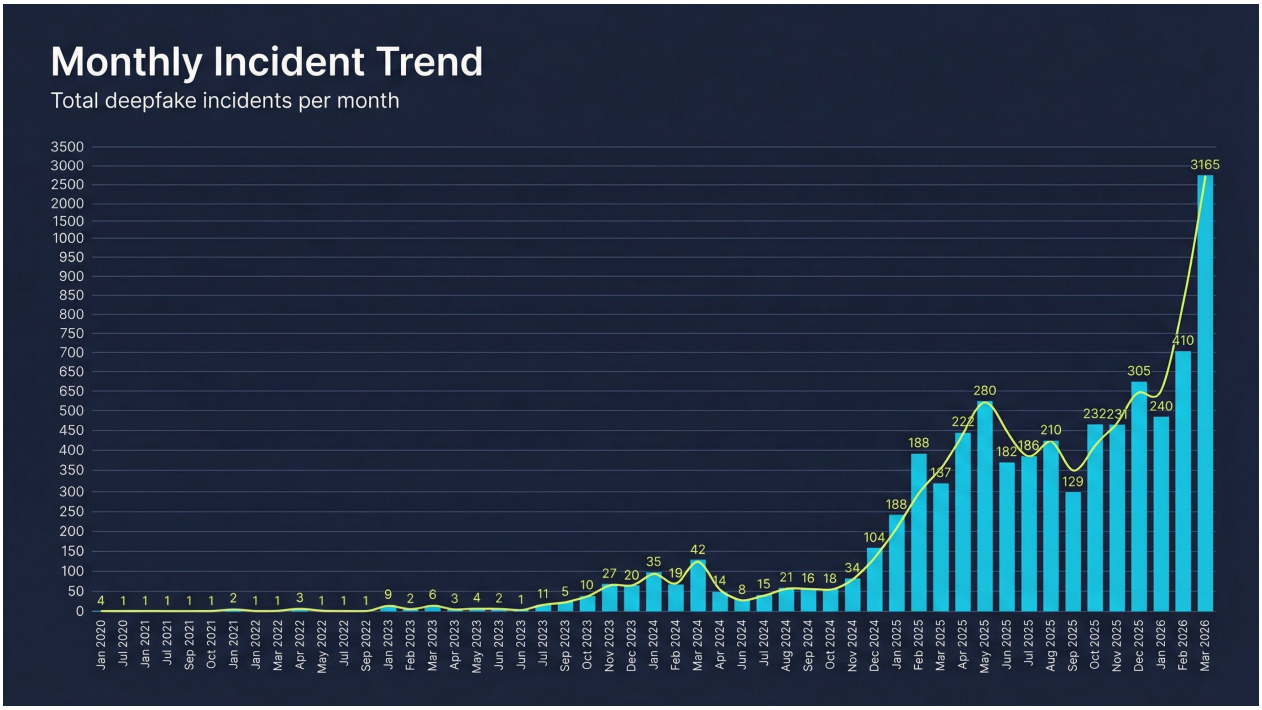


Figure 3: Monthly incident trend – total deepfake incidents logged per month.

The dominance of AI-generated video is largely underpinned by the evolution of Generative Adversarial Networks (GANs) and sophisticated diffusion models. These technologies power high-fidelity face-swap operations and full-face synthesis, techniques predominantly leveraged in extensive political disinformation campaigns and elaborate celebrity-endorsed investment frauds. Concurrently, the **10.5%** share of voice cloning maps directly to advancements in Text-to-Speech (TTS) pipelines and voice-conversion technologies. These audio carriers require increasingly minimal zero-shot training data—often just seconds of a target's voice—enabling precise, synchronous Business Email Compromise (BEC) attacks and vishing operations. Synthetic still images, relying on static diffusion models, remain a highly efficient vector for generating fraudulent documentation and non-consensual explicit material, operating as low-overhead, high-impact payloads.

Tracing the temporal evolution of these attack vectors reveals critical inflection points over the multi-year observation period. In the early stages of 2020, synthetic still images and rudimentary video face-swaps exhibited a fluctuating percentage share, primarily constrained by the heavy computational overhead required for generation. However, a notable shift occurred in late 2022 and early 2023, coinciding with the widespread release of optimized, consumer-grade generative frameworks. During these peak months, the percentage share of mixed media formats began to surge, indicating a shift toward multi-modal deception. Voice cloning also saw an accelerating temporal trajectory in late 2023 through 2024, transitioning from an experimental technique to a mainstream operational tool for enterprise fraud, as real-time latency barriers were systematically broken down by threat actors. Figure 4 illustrates the carrier mix evolution, showing the monthly percentage share of each deepfake modality.

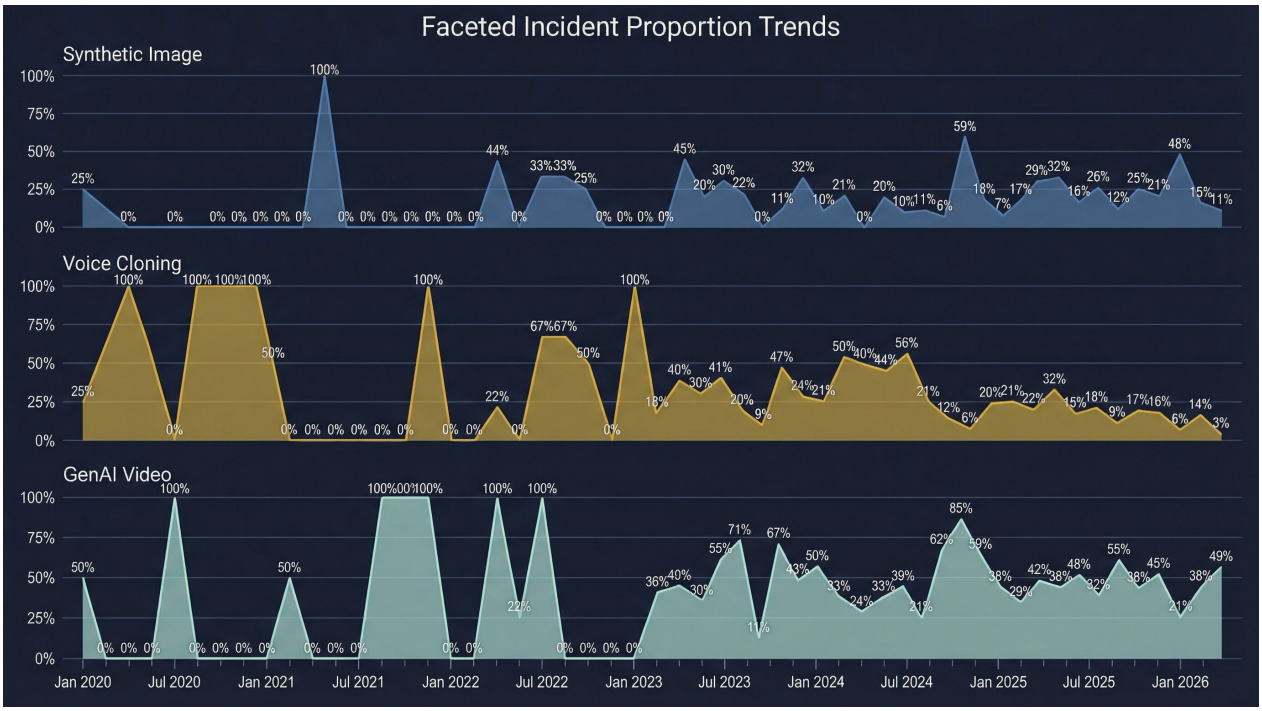


Figure 4: Carrier mix evolution – monthly percentage share of each deepfake modality.

Crucially, the evolution of these technical modalities is now inextricably linked to the escalating threat against identity infrastructure. Automated camera injection tools and virtual webcam drivers are increasingly paired with real-time video and audio generation to execute sophisticated biometric verification and Know Your Customer (KYC) bypass attacks. Attackers bypass physical sensors entirely, injecting synthetic video streams directly into the application layer of banking and enterprise onboarding portals. This evolution represents a shift from cognitive deception–tricking a human observer–to algorithmic deception, where the technical carriers are specifically engineered to defeat liveness detection algorithms and mathematical verification thresholds.

4 Cross-Analysis: Threat Vectors & Targets

A multidimensional examination of the synthetic media landscape reveals complex, underlying correlations between threat vectors, technical modalities, targeted entities, and platform distribution mechanisms. Analyzing the interactions across these variables is paramount for understanding the operational calculus of threat actors. Based on our observations spanning from early 2020 to early 2026, the intersectional data illustrates that attackers do not select their tools or platforms arbitrarily; rather, they orchestrate highly optimized campaigns where the carrier method and distribution channel are precisely tailored to the intended psychological or financial impact.

The correlation between political manipulation (**24.6%** of classified categories) and high-velocity social media environments is exceptionally pronounced. Political subversion operations index heavily against the ecosystem's most dominant technical carriers: video (which accounts for **45.6%** of all incidents globally) and mixed media formats (**25.2%**). The strategic rationale is evident: political disinformation relies on the rapid, viral spread of emotionally resonant visual narratives before fact-checking mechanisms can intervene. Consequently, these multi-sensory formats find their optimal operational theater on platforms like X (**61.3%**) and YouTube (**13.4%**). The algorithmic architectures of these platforms, which prioritize engagement and retention, inadvertently serve as force multipliers for synthetic video content. A deep analysis of these intersecting clusters indicates that state-aligned actors and decentralized provocateurs actively engineer political deepfakes to exploit the specific vertical-scrolling, short-attention-span dynamics of these networks. Figure 5 depicts the incident density heatmap correlating attack categories with platform distribution. To properly interpret the incident density heatmap in Figure 5, the data should be read horizontally from left to right. Each row represents a distinct incident category as the primary subject. The percentages across a given row illustrate how that specific threat category is distributed across various social media platforms. For instance, this row-wise view clarifies which platforms are uniquely favored for political manipulation versus celebrity misuse.

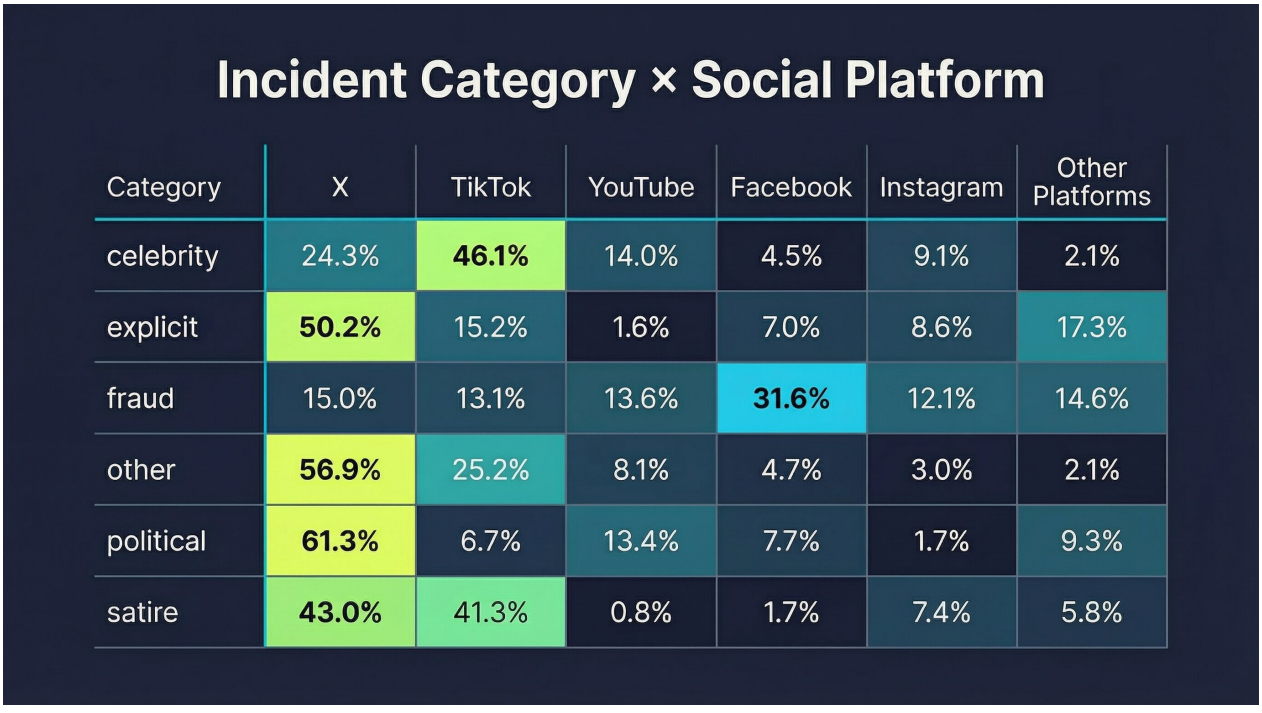


Figure 5: Incident category × social platform – incident density heatmap.

Conversely, the deployment of synthetic media for enterprise fraud (**20.1%**) exhibits a diverging structural pattern. While fraud campaigns certainly utilize video for broader consumer-facing investment scams (often heavily overlapping with the **9.3%** celebrity misuse category), targeted corporate fraud and Business Email Compromise (BEC) demonstrate a strong intersection with audio modalities (**10.5%**). Voice cloning technologies, requiring significantly less source material to

train than high-fidelity video models, are increasingly deployed in synchronous attacks against financial departments and executive teams. These audio-centric fraud vectors rarely propagate through public-facing platforms like Instagram or TikTok. Instead, they bypass social media entirely, infiltrating direct telephonic communications or encrypted messaging channels like Telegram (**1.4%**), where the absence of visual context forces the victim to rely solely on compromised vocal verification.Figure 6 details the incident density heatmap of deepfake types deployed across various target sectors. The data in Figure 6 is designed to be read vertically by column. Each column focuses on a specific targeted sector as the primary subject. The percentages read from top to bottom reveal the breakdown of technical modalities (such as video, image, or audio) deployed against that particular target group. This vertical analysis highlights how threat actors shift their technical carriers depending on the exact profile of the victim.

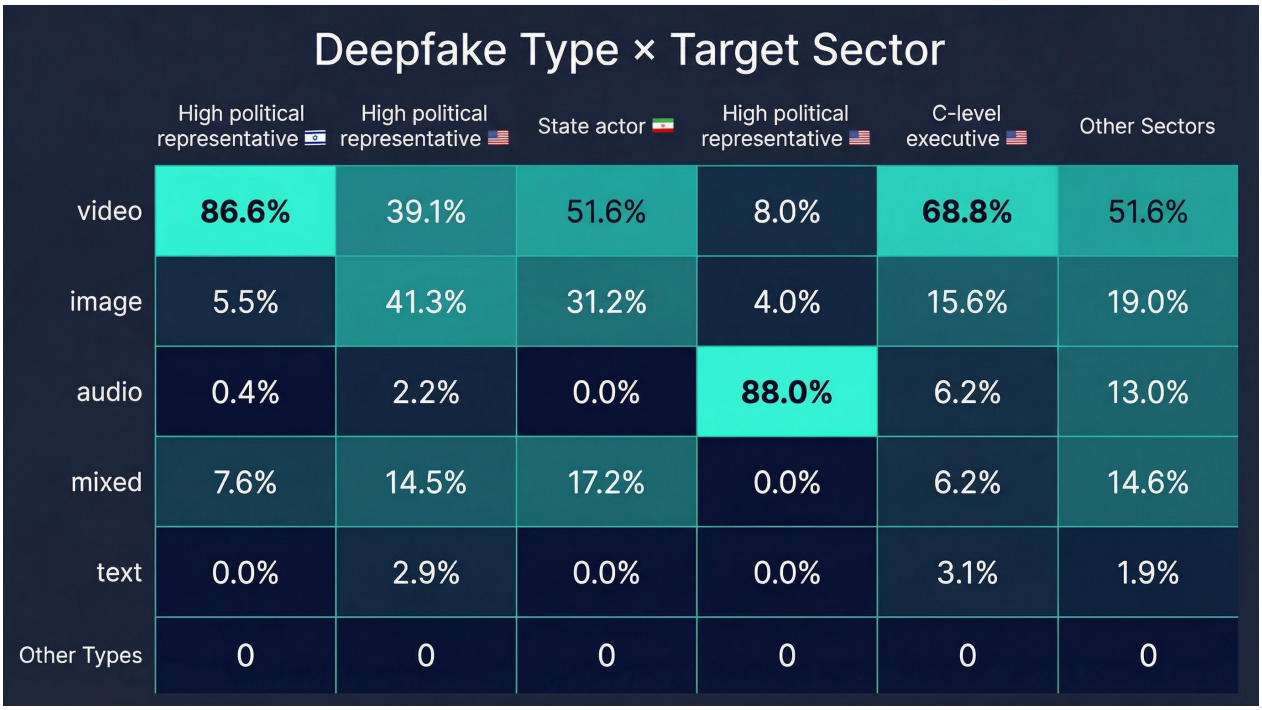


Figure 6: Deepfake type × target sector – incident density heatmap.

Furthermore, the data indicates a critical interaction between the geographical footprint and platform targeting. Threats originating or culminating in the United States (**46.9%**) frequently exhibit sophisticated multi-platform orchestration. For instance, a campaign might utilize a synthetic still image (**17.4%**) seeded on X (which captures **51.2%** of social media distribution) to establish a false narrative, which is subsequently validated by artificially generated text (**1.3%**) and amplified via algorithmic surges on TikTok. This cross-pollination of modalities and platforms creates a resilient, interlocking web of synthetic evidence that overwhelms traditional, single-platform content moderation strategies.

The implications of these multi-variable correlations are severe. Threat actors are no longer relying on isolated, monolithic deepfakes. They are deploying composite attacks where the technical modality is dynamically adjusted based on the target's specific vulnerabilities and the chosen platform's distinct algorithmic behavior. As attackers continue to refine this matrix of variables, defensive postures must evolve beyond simple media forensics to encompass holistic, cross-platform behavioral analysis.

5 12-Month Threat Forecast

Based on the temporal evolution and technical trajectory derived from our observations, the next 12 months will mark a critical transition phase for synthetic media operations. The growth vectors for the upcoming year are projected to shift from asynchronous, pre-rendered media campaigns toward synchronous, real-time exploitation. The foundational technologies enabling this escalation—real-time face-swap engines, low-latency live voice cloning APIs, automated document synthesis kits, and sophisticated camera injection frameworks—have moved beyond proof-of-concept and are currently being commoditized within cybercriminal marketplaces.

We assess three systemic risk scenarios that organizations must prepare for over the next 12-month horizon:

Firstly, KYC Bypass at Scale carries a high likelihood and critical impact. As financial institutions increasingly rely on remote identity verification, threat actors will industrialize the use of camera injection frameworks paired with real-time biometric manipulation. We project an escalation in attacks successfully defeating standard liveness detection, leading to the mass creation of synthetic mule accounts and the subsequent fracturing of digital trust within retail banking sectors.

Secondly, the Industrialisation of BEC Voice Fraud represents a high likelihood and severe impact scenario. Extrapolating the current growth in audio modalities, financially motivated syndicates will likely automate voice cloning at scale. By integrating live voice APIs with compromised enterprise communication directories, attackers will execute synchronous, multi-lingual vishing campaigns targeting C-suite executives and finance departments, drastically increasing the financial yield of traditional social engineering.

Thirdly, a Disinformation Surge is projected with a moderate-to-high likelihood and structural impact. Leveraging mixed media formats and highly targeted algorithmic deployment, state-aligned actors will orchestrate localized, hyper-specific cognitive disruption campaigns. These operations will utilize synthetic video and audio to fabricate events in real-time, deliberately polluting the information environment during critical geopolitical windows.

As an early-warning signal, the industry must rigorously monitor the rapid emergence of autonomous, multi-modal synthetic campaigns. We are observing the nascent integration of large language models governing both the conversational logic and real-time generation of synthetic avatars. In the near future, threat actors will deploy fully autonomous agents capable of conducting interactive, highly persuasive video and audio engagements with human targets, fundamentally redefining the scale and velocity of social engineering.


GEOGRAPHY REPORT
GLOBAL INCIDENT ANALYSIS & HOTSPOT VISUALIZATION
GEOGRAPHY
USA
CHINA
INDIA
BRAZIL
Global Deepfake Spread

6 Geopolitical Footprint

An analysis of the geopolitical footprint derived from our observations highlights a highly concentrated and asymmetrical threat environment. The geographic distribution of deepfake incidents, as depicted in Figure 7, heavily impacts major democratic and economic hubs, with the United States absorbing the dominant share at **46.9%**. The United Kingdom follows at **8.2%**, India at **7.2%**, and Israel at **6.6%**. Notably, incidents of unknown or heavily obfuscated origins account for **4.5%** of the distribution, reflecting the sophisticated operational security measures employed by advanced threat actors.

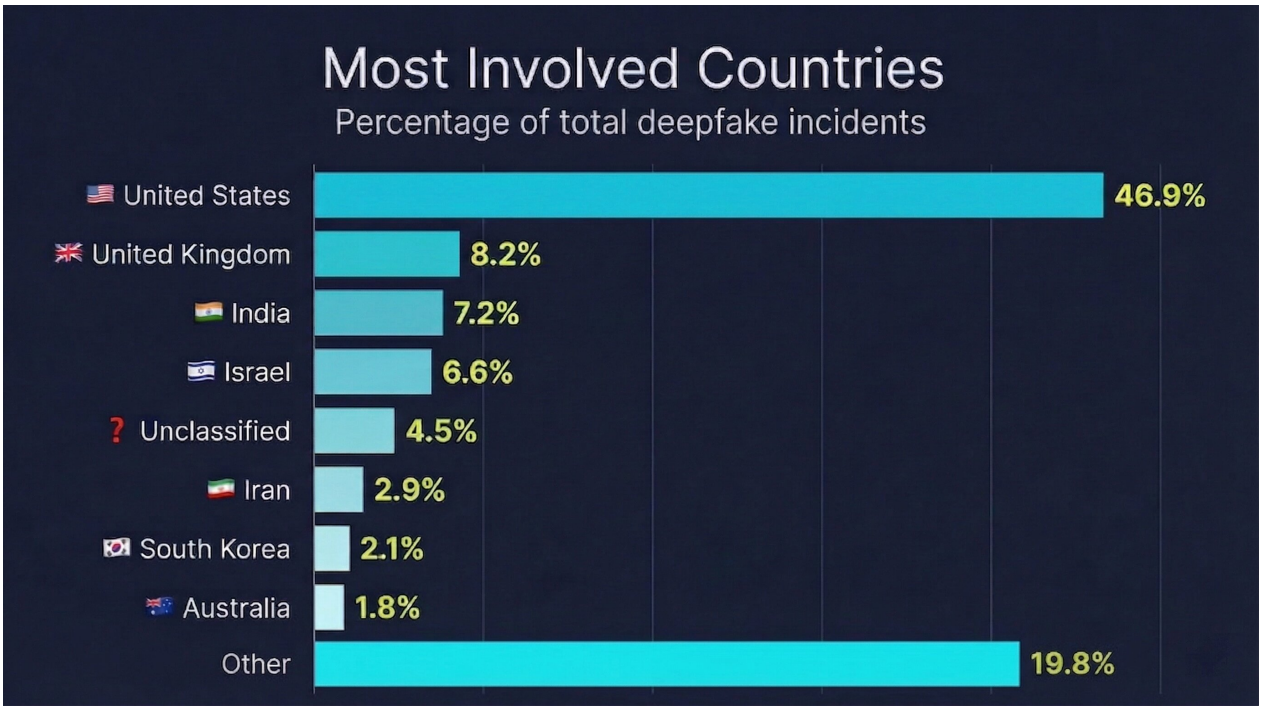


Figure 7: Most affected countries – percentage share of incidents by geography.

This geographic clustering reveals severe cross-border threat flows and distinct asymmetries between regions of origin and ultimate target destinations. Threat actors—often operating from jurisdictions with permissive cybercrime environments or tacit state sponsorship—systematically target the populations and institutional infrastructures of the US, UK, and allied nations. The objective is to exploit the open digital ecosystems and high algorithmic velocity inherent to Western media consumption. Consequently, target nations bear the disproportionate burden of societal disruption and financial fraud, while the technical infrastructure enabling the attacks is distributed globally, as illustrated in Figure 8.

These localized threat percentages map directly onto the evolving global regulatory environment, which is fracturing into distinct legislative approaches. In Europe, the EU AI Act aims to impose stringent transparency and provenance requirements on generative models, though the UK (**8.2%**) is presently navigating a more agile, context-driven regulatory framework post-Brexit. In the United States (**46.9%**), the response is largely characterized by the US Executive Order on AI and fragmented state-level legislation, focusing on federal procurement standards and sector-specific guidelines. Meanwhile, in APAC regions like India (**7.2%**), emerging frameworks are rapidly attempting to balance sweeping digital public infrastructure protection with aggressive countermeasures against political disinformation. These disparate regulatory postures create jurisdictional arbitrages that threat actors actively exploit, underscoring the necessity for harmonized, cross-border intelligence sharing and standardized cryptographic provenance.

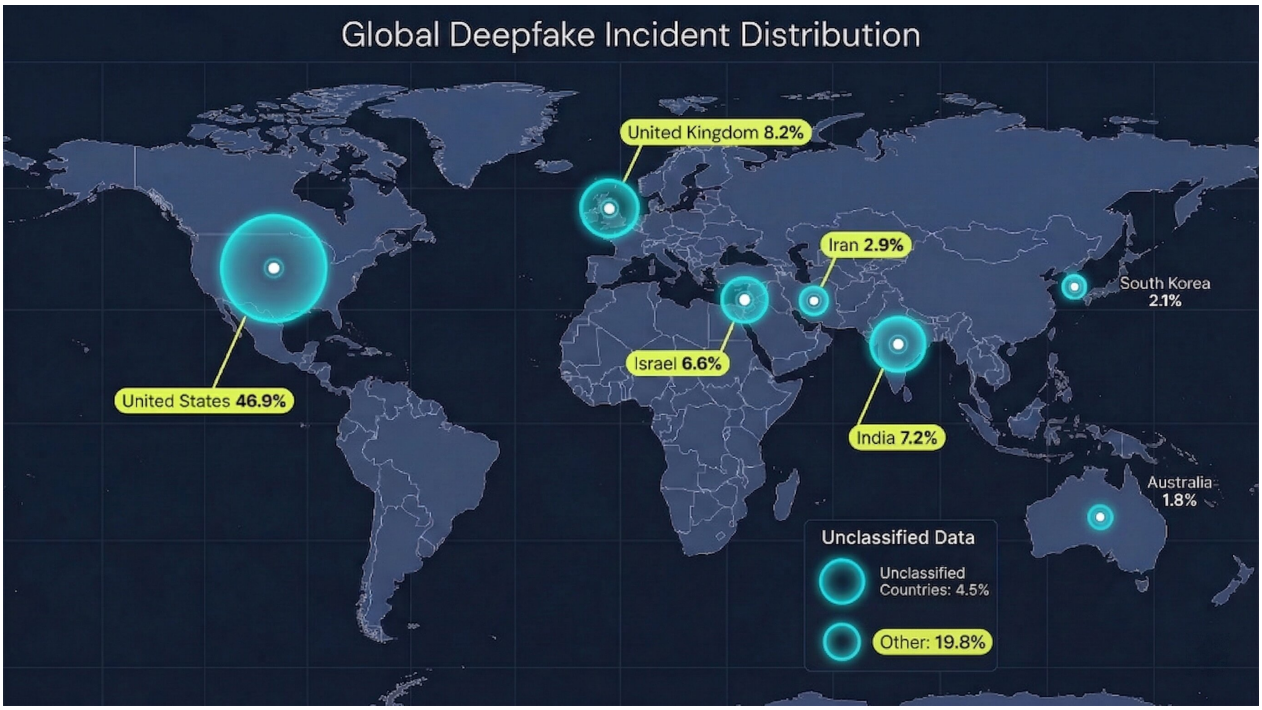


Figure 8: Global deepfake incident distribution – proportional geographic footprint.

7 Victim Profile & Targeted Sectors

The profile of victims and targeted sectors within our observations reveals a bifurcated strategy employed by threat actors, simultaneously targeting high-value individual identities for maximum psychological impact and systematically probing broader institutional structures for financial extortion.

High-value identity targets command a disproportionate share of visibility within the threat landscape. Political leaders and prominent public figures are subjected to relentless synthetic impersonation to manipulate geopolitical narratives. For example, a senior Israeli political leader was the targeted subject in **5.4%** of all observed incidents, underscoring the weaponization of synthetic media in acute regional conflicts. Similarly, a prominent US political figure accounted for **3.1%** of incidents, reflecting persistent domestic election interference and polarization efforts. Additionally, state entities themselves are not immune to targeting; a Middle Eastern state entity was implicated as the subject of manipulation in **1.4%** of incidents.

Beyond these high-profile individuals, the broader population remains persistently vulnerable. The general public comprises a combined **1.7%** of targeted incidents, primarily falling victim to mass-distributed financial scams, localized non-consensual imagery, and decentralized identity theft. While the percentage share affecting everyday citizens may appear numerically smaller against global leaders, the severe financial and reputational impactat the individual level is profound.

Crucially, an analysis of percentage-point shifts indicates a rapid pivoting of attack strategies toward the enterprise sector. The fastest-growing victim cohorts are C-level executives, finance directors, and institutional leaders. This shift marks the industrialization of synthetic media from a tool of public disinformation to a precise instrument for corporate financial fraud. Threat actors are recognizing that impersonating an enterprise leader yields direct, lucrative monetization through Business Email Compromise (BEC) and authorized push payment fraud, circumventing traditional network security perimeters by exploiting the human element within financial workflows.

8 Social Platform Exploitation

Platform exploitation is the fundamental distribution engine for the deepfake threat ecosystem. Based on our observations, the distribution of synthetic media incidents across social architectures demonstrates a clear hierarchy of threat actor preference. X accounts for a commanding **51.2%** of incident propagation, followed by TikTok at **21.1%**, YouTube at **10.0%**, Facebook at **8.2%**, Instagram at **5.0%**, Telegram at **1.4%**, WhatsApp at **1.1%**, and other platforms at **2.0%**. Unclassified or unknown platform origins account for the remainder, reflecting instances of cross-platform seeding or closed-network distribution. Figure 9 provides an overview of this platform distribution, including individual breakdowns for social networks and news outlets.

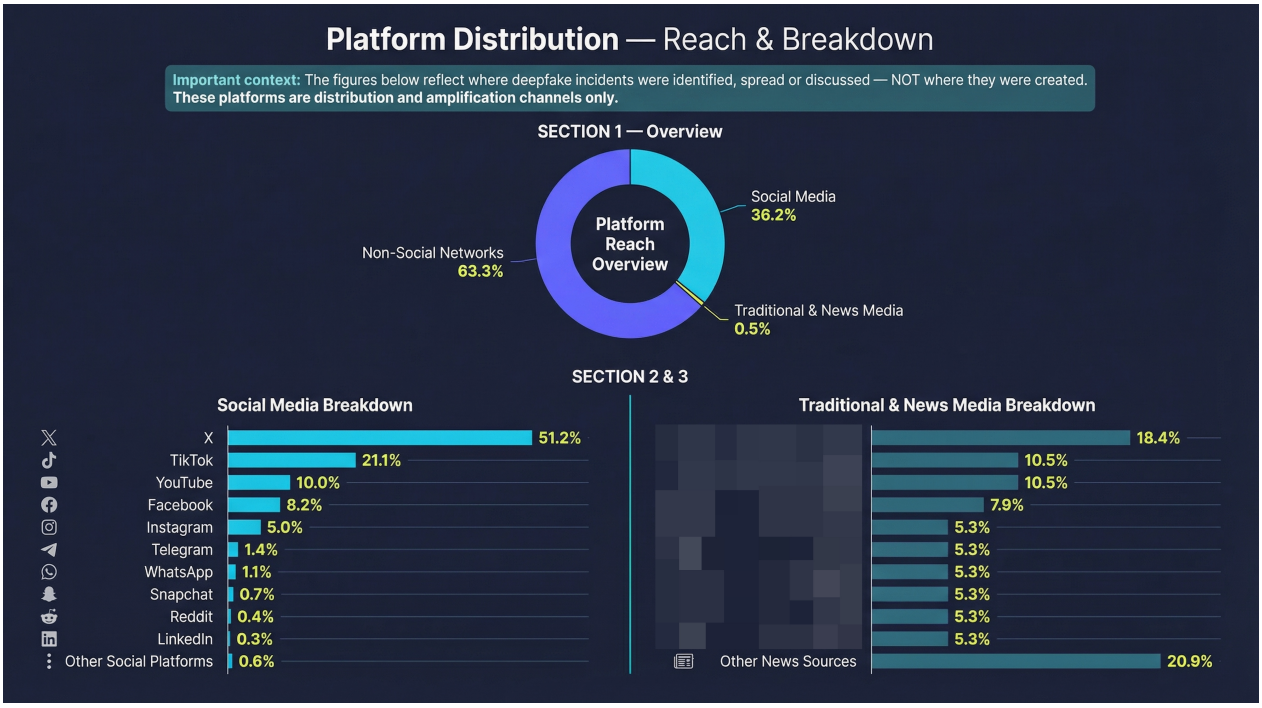


Figure 9: Platform distribution – overview of social vs. traditional media reach with individual breakdowns for social networks and news outlets. Note: platforms are distribution channels, not content generators.

This distribution directly exposes the structural vulnerabilities inherent to these platforms. X's massive **51.2%** share is a direct consequence of its hyper-optimized algorithmic amplification dynamics, which prioritize immediate engagement and rapid content velocity over source authentication. Threat actors exploit this velocity, recognizing that synthetic video content can achieve millions of views and shape public perception long before content moderation workflows can successfully identify and action the media. YouTube and Facebook share similar vulnerabilities regarding user scale and weak proactive moderation at the perimeter, allowing fraudulent investment schemes and synthetic political narratives to incubate within vast, segmented audience pools. Telegram's **1.4%** share, while comparatively smaller, is structurally vital for threat actors; its optionally encrypted, pseudonymous environment serves as both a distribution channel for high-harm audio/video payloads and a marketplace for the trading of synthetic media generation tools.

Social media platforms are no longer passive intermediaries; they operate as the primary epistemological layer for global populations. Consequently, they hold a structural responsibility to integrate deeply into the trust architecture required to defend truth. The era of reactive takedowns is structurally inadequate. Fast-evolving global regulations are correctly demanding heightened platform accountability. The mandate is clear: platforms must transition toward algorithmic transparency, cryptographic provenance integration, and proactive, mathematically rigorous detection at the point of upload, ensuring that synthetic deception is neutralized before it achieves algorithmic virality.

9 Media Amplification & Journalistic Coverage

Data Anonymization Strategy

The decision to anonymize newspaper sources while maintaining the identity of social network platforms stems from the fundamental difference in their **content authorship**. Social media content is generated by a vast, heterogeneous pool of individual users; therefore, the platform itself does not represent a single editorial voice or a deliberate corporate stance. In contrast, newspaper articles reflect the specific **editorial line** and strategic direction of a publisher. Since our study aims to analyze the landscape of media discourse without casting judgment or establishing direct comparisons between specific outlets, we chose to anonymize these sources. This approach allows us to focus the analysis on sector-wide trends, highlighting how certain publications show a higher sensitivity to the "deepfake" phenomenon during the observation period, while others remain more peripheral, all without introducing institutional bias into our findings.

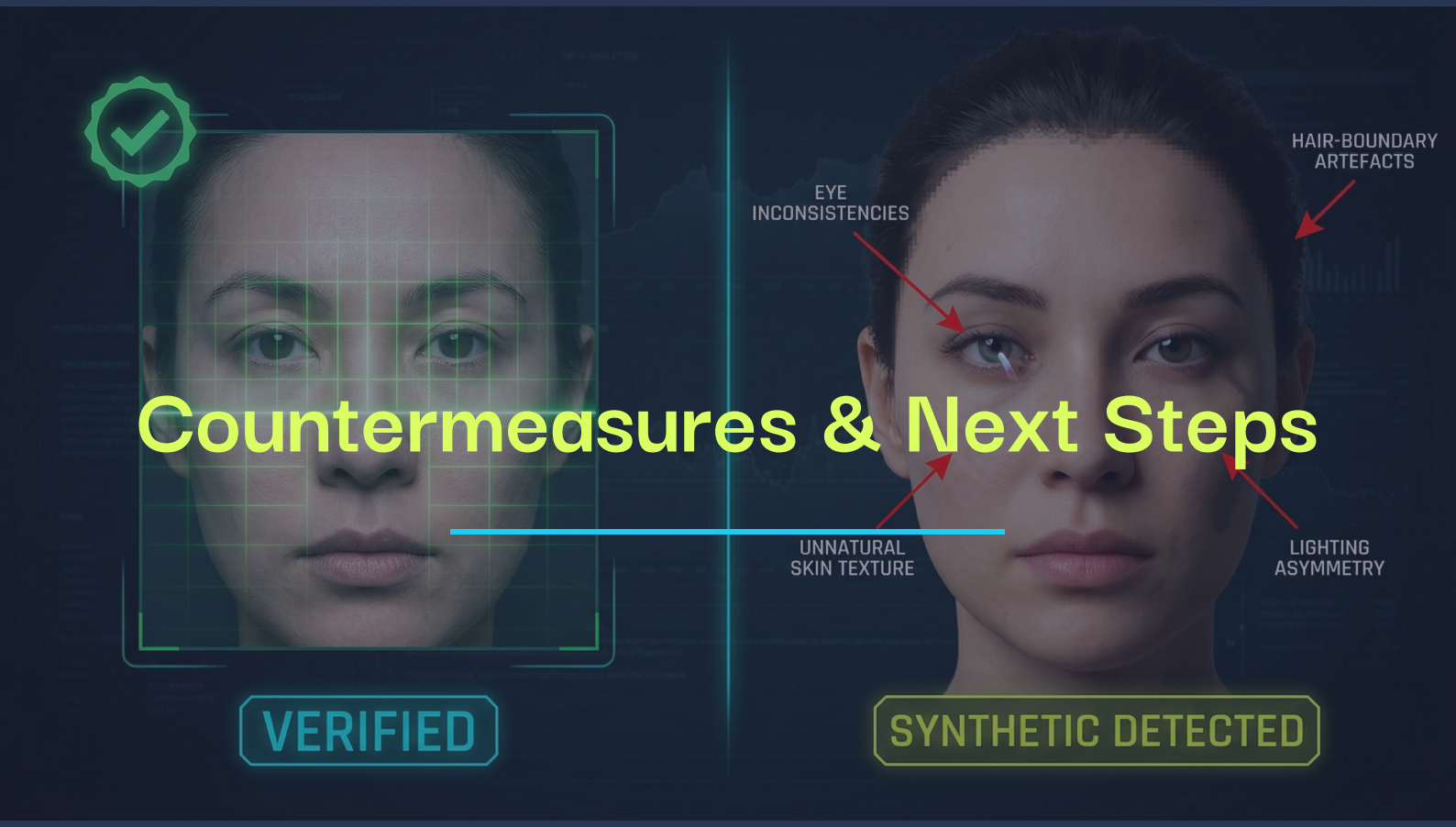
Journalism remains one of the most consequential defensive layers in the synthetic media ecosystem. News organisations, fact-checkers, broadcast bodies, and specialist reporters continue to play a vital role in identifying, verifying, and contextualising manipulated content – often before it achieves mainstream acceptance as authentic.

Our observations reflect coverage distributed across a diverse mix of outlet types: mainstream national press, broadcast organisations, technology and specialist publications, wire services, and regional-national titles. This breadth should not be read as evidence of media complicity in synthetic media propagation. On the contrary, it reflects the scale and seriousness of the public-interest response – outlets across the spectrum surfacing incidents, scrutinising claims, and helping audiences navigate the increasingly blurred boundary between authentic and manufactured content.

No single outlet type dominates. Broadcast and mainstream national press account for the largest combined share, followed by specialist technology journalism and wire services – a distribution that underscores how deepfake-related coverage has moved beyond niche reporting into mainstream editorial agendas.

The structural challenge, however, is intensifying. Synthetic media is faster, cheaper, and more persuasive than at any previous point, while editorial teams operate under sustained time pressure in a high-velocity information environment. Threat actors are not pursuing virality alone – they are pursuing legitimacy, engineering manipulated content to survive long enough to enter mainstream discourse and shape wider narratives before correction can occur.

The key conclusion is not that journalism is failing. It is that journalism alone cannot absorb the full verification burden once synthetic content begins to scale. Editorial scrutiny remains essential – but it cannot be the only safeguard. Real-time detection, authenticity infrastructure, and stronger platform accountability are no longer optional supplements to journalistic rigour. They are structural necessities.



Countermeasures & Next Steps

10 Strategic Countermeasures & Mitigation Roadmap

Securing the integrity of digital infrastructure against the proliferation of synthetic media requires a coordinated, multi-layered defensive posture. The following mitigation roadmap categorizes strategic countermeasures by stakeholder segment, derived directly from our operational observations of deepfake TTPs.

KYC & Identity Verification Providers

Identity verification architectures are the primary targets for automated injection and biometric spoofing. Upgrading these defenses requires immediate technical intervention.

1. **Deploy Camera Injection Defences (Immediate):** Implement protocol-level validation to ensure media streams originate from authenticated physical hardware sensors, rejecting virtually routed camera feeds. *Success Indicator:* A **95%** reduction in successful onboarding attempts utilizing known virtual webcam drivers and emulator environments.
2. **Calibrate Liveness Detection Thresholds (Immediate):** Upgrade active and passive liveness checks to dynamically detect temporal inconsistencies, rendering artifacts, and micro-expression failures unique to GAN-generated synthetic video. *Success Indicator:* Zero-day detection of face-swap modalities during synchronous identity verification challenges.
3. **Integrate Multi-Modal Biometric Checks (12-Month):** Require synchronized, cross-modal verification (e.g., simultaneous vocal and facial analysis) to mathematically correlate audio-visual streams, vastly increasing the computational cost for attackers. *Success Indicator:* A quantifiable increase in the required attacker time-to-compromise, leading to higher abandonment rates.

Enterprise & Organisational Risk Teams

As threat actors pivot heavily toward corporate extortion, organizational resilience must transition from basic awareness to structural defense.

1. **Implement Vishing/BEC Deepfake Playbooks (12-Month):** Author and enforce out-of-band verification protocols for high-risk financial transactions, requiring multi-channel authentication independent of voice or video calls. *Success Indicator:* Zero unauthorized financial transfers executed solely on the basis of synchronous executive communication.
2. **Integrate Multi-Modal Deepfake and GenAI Detection (Immediate):** Procure and deploy deep-learning-based media forensics tools, such as the IdentifAI detection engine, across corporate systems to automatically flag manipulated audio, video, and image synthesis. These systems should leverage structural deconstruction to identify inconsistencies in media that bypass standard liveness checks. *Success Indicator:* The automated quarantine of at least **90%** of known synthetic payloads before they reach the end-user.
3. **Execute Targeted Staff Awareness Simulations (12-Month):** Conduct controlled, simulated synthetic voice attacks against finance and HR personnel to practically assess human vulnerability and response times. *Success Indicator:* A **50%** improvement in employee reporting velocity during simulated voice-cloning vishing exercises.

Policymakers & Regulators

Legislative and standards-setting bodies must establish the foundational frameworks that force transparency and accountability across the digital ecosystem.

1. **Advance Digital Provenance Standards (Strategic):** Mandate the integration of cryptographic provenance frameworks, such as C2PA, across consumer hardware and social distribution platforms to guarantee media origin transparency. *Success Indicator:* Legislative adoption requiring mandatory watermarking and metadata preservation for commercially available generative models.

2. **Clarify Platform Liability Frameworks (Strategic):** Revise existing safe harbor protections to ensure social media platforms bear structural liability for the algorithmic amplification of non-consensual and fraudulent synthetic media. *Success Indicator:* The measurable reallocation of platform engineering resources toward proactive, pre-upload detection capabilities.
3. **Establish Cross-Border Incident Registers (Strategic):** Fund and deploy centralized, international clearinghouses for deepfake threat intelligence, enabling rapid sharing of IoCs (Indicators of Compromise) between allied nations. *Success Indicator:* The successful, real-time syndication of synthetic threat signatures across major democratic cyber-defense agencies.

11 Examples

The rapid advancement of artificial intelligence and deepfake technology has introduced new challenges across various sectors. The following examples highlight recent incidents involving AI manipulation and synthetic media in politics, celebrity culture, and financial fraud.

Politics: AI Manipulation Theories

In March 2026, viral rumors circulated across social media claiming that Israeli Prime Minister Benjamin Netanyahu had died. These rumors were heavily fueled by a video address in which a peculiar camera angle created an optical illusion, making it appear as though Netanyahu had six fingers on one hand. Conspiracy theorists quickly seized on this frame, arguing that the footage was an AI-generated deepfake meant to cover up his death amid heightened geopolitical tensions. Fact-checkers soon debunked the claim, and Netanyahu subsequently released follow-up videos directly addressing the public and displaying his hand to prove the initial video was not an AI fabrication.

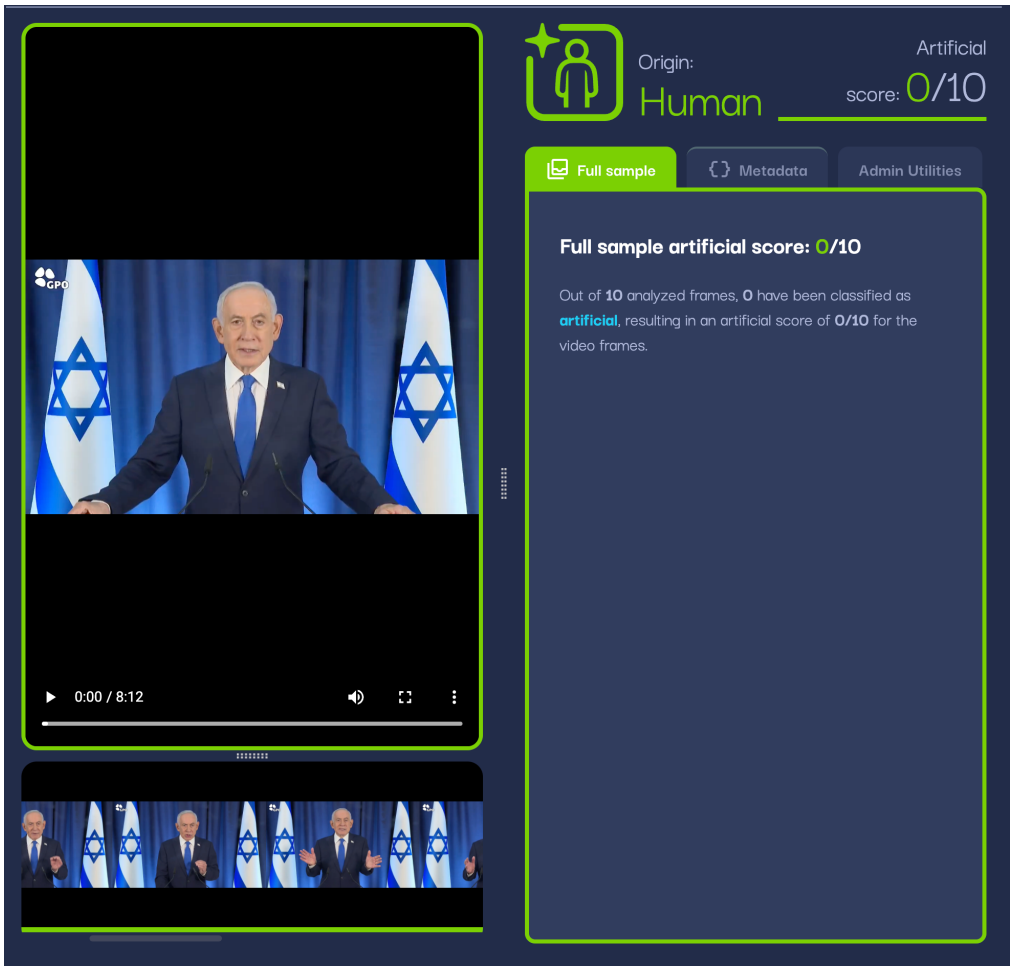


Figure 10: identifAI analysis of one of the alleged deepfake videos from social media.

Celebrities: Unauthorized Likeness and Deepfakes

Public figures are increasingly falling victim to unauthorized digital replication. A notable incident reported by the BBC involved actress Scarlett Johansson, who publicly warned against the "misuse of AI" after a deceptive deepfake video surfaced featuring a fabricated version of Kanye West. The controversy underscores the vulnerability of celebrities to AI identity theft and highlights the urgent need for robust regulatory frameworks to protect individuals from having their

likenesses hijacked for malicious, defamatory, or non-consensual purposes.

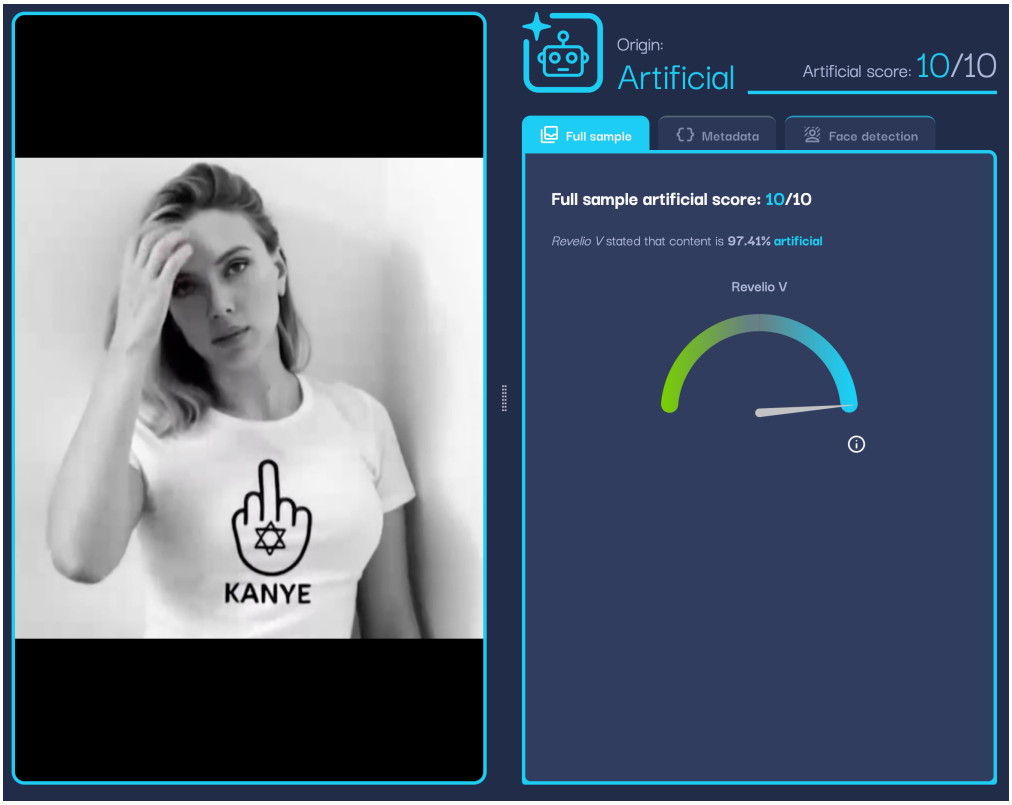


Figure 11: identifAI analysis of the deepfake video involving Scarlett Johansson.

Fraud: Sophisticated Crypto Heists

Cybercriminals are now integrating real-time deepfakes into complex social engineering attacks. In a campaign discovered in early 2026, a North Korean state-backed hacking group (UNC1069) targeted financial technology and cryptocurrency firms. The attackers compromised the Telegram account of a crypto executive and invited victims to a fake Zoom meeting. Once on the call, targets were presented with a convincing deepfake video of the executive. The hackers then faked an "audio issue" and instructed the victims to run a command to fix it (a "ClickFix" attack). This deceptive tactic tricked the victims into installing macOS malware, allowing the hackers to harvest credentials, breach the network, and steal cryptocurrency.

12 Methodology

The intelligence and statistical models presented in this report are the result of a rigorous incident harvesting and classification pipeline. Our methodology is designed to capture a comprehensive global observation of deepfake incidents, operating over the continuous reporting period from January 2020 through March 2026. To maintain absolute objectivity and operational transparency, this analysis strictly excludes unverified proprietary telemetry, relying entirely on the aggregation of public indicators.

An incident is classified and ingested into our observation framework solely if it represents a case of a deepfake being deployed in the real world, substantiated by English-based media reporting globally or through explicit mentions of synthetic media operations on social media platforms. Once identified, the unstructured data is processed through an advanced LLM-powered classification architecture. This analytical layer performs entity extraction, modality identification, and thematic categorization, effectively curating the raw intelligence into a structured attack taxonomy.

By applying LLM-powered analysis, we isolate the specific Tactics, Techniques, and Procedures (TTP) utilized by threat actors, mapping complex variables such as platform propagation and geographic origin. To ensure precise structural insights, all aggregated data is subjected to strict percentage-based statistical modeling. This percentage-only approach neutralizes the volume-based biases inherent in raw data collection, allowing us to accurately track temporal shifts, assess the evolving operational priorities of threat actors, and calculate the exact proportional distribution of risks across the synthetic media ecosystem.